Identifying Telescope Usage in Astrophysics Publications: A Machine Learning Framework for Institutional Research Management at Observatories

Vicente Amado Olivo[1] , Wolfgang Kerzendorf[1,2] , Brian Cherinka[3] , Joshua V. Shields[2] , Annie Didier[4], and Katharina von der Wense[5,6]

[1] Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI 48824, USA; amadovic@msu.edu
[2] Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA
[3] Space Telescope Science Institute, Baltimore, MD 21218, USA
[4] Outrider, Golden, CO 80403, USA
[5] Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309, USA
[6] Institute of Computer Science, Johannes Gutenberg University Mainz, 55128 Mainz, Germany

Abstract

Large scientific institutions, such as the Space Telescope Science Institute, track the usage of their facilities to understand the needs of the research community. Astrophysicists incorporate facility usage data into their scientific publications, embedding this information in plain text. Traditional automatic search queries prove unreliable for accurate tracking due to the misidentification of facility names in plain text. As automatic search queries fail, researchers are required to manually classify publications for facility usage, which consumes valuable research time. In this work, we introduce a machine learning classification framework for the automatic identification of facility usage of observation sections in astrophysics publications. Our framework identifies sentences containing telescope mission keywords (e.g., Kepler and TESS) in each publication. Subsequently, the identified sentences are transformed using term frequency–inverse document frequency and classified with a support vector machine. The classification framework leverages the context surrounding the identified telescope mission keywords to provide relevant information to the classifier. The framework successfully classifies the usage of MAST-hosted missions with a 92.9% accuracy. Furthermore, our framework demonstrates robustness when compared to other approaches, considering common metrics and computational complexity. The framework's interpretability makes it adaptable for use across observatories and other scientific facilities worldwide.

Unified Astronomy Thesaurus concepts: Support vector machine (1936); Random Forests (1935); Classification (1907); Open source software (1866)

1. Introduction

The global astronomical community identifies the scientific output of telescope facilities, which is essential for understanding the needs of the community, effectively allocating resources, and planning current and future facility management. A primary metric for measuring this scientific output is tracking the usage of telescope facilities in scientific publications (see, e.g., U. Grothkopf et al. 2018). However, simple programmatic attempts at identifying telescope usage based on the appearance of mission keywords fail because the descriptions of telescope usage are embedded in an unstructured way throughout the plain text of publications. Consequently, for now, researchers are required to manually identify publications and leverage specific subject knowledge needed for accurate identification.

The exponential growth of the astronomical literature requires researchers to continually spend additional time manually classifying publications, a mundane task that diverts researchers from their primary research (see, for example, W. E. Kerzendorf 2019; L. Bornmann et al. 2021). At the Space Telescope Science Institute (STScI), a small team of researchers queries the Smithsonian Astrophysical Observatory/NASA Astrophysics Data System (NASA/ADS)[7] each month, dedicating part of their time to identifying and classifying as many relevant publications as possible, alongside their other responsibilities. For example, from 2021 to 2023, on average, ~550 new publications per month were identified for classification. Given workforce capacity and resource limitations, on average, 14% of these publications remain unclassified, contributing to a growing backlog (B. Cherinka 2024, private communication). The accurate and scalable tracking of telescope usage can potentially be met by automatic text classification.

Simple automatic solutions, such as full-text searches, cannot accurately distinguish the nuances that indicate telescope usage in astrophysics publications (T. Barnett et al. 2009). Facility names (e.g., FUSE, Copernicus, and Kepler) can be misidentified owing to their lack of uniqueness and potential overlap with common words (see Figure 1). While the lack of structure in the plain text of publications requires initially searching for telescope names, accurate classification requires leveraging the context to identify telescope usage.

Text classification is a principle task in the field of natural language processing (NLP), with various recent advancements with the specific aim of allowing for the inclusion of domain knowledge and feedback processes (e.g., fine-tuning; I. Beltagy et al. 2019). Accordingly, much research has been directed at

---

[7] https://ui.adsabs.harvard.edu/

Figure 1. The filtered data set comprises sentences that include MAST-hosted mission keywords. For example, when filtering publications for the space telescope "Kepler," we identify three sentences shown in these two publications. Contextual cues guide the reader or classifier to distinguish that, on the left, one publication references the "Kepler" space telescope and is a MAST publication, while on the right, the other publication discusses "Keplerian physics" and the research does not utilize the Kepler space telescope.

presenting efficient, generalizable, and accurate methodologies with more applications appearing recently in the scientific domain (M. Khadhraoui et al. 2022; D. Farshid & M. Dastani 2023). As simple automatic solutions fail, the advancements of computational techniques, like machine learning, are leveraged for the task of identifying telescope or data usage from publications (T. X. Chen et al. 2022, hereafter C22).

In the domain of astrophysics, C22 presents a classification framework trained to categorize publications relevant to NASA/IPAC Extragalactic Database (NED) data products. Utilizing the Stanford NLP classifier, C22 identifies NED-relevant publications. Nevertheless, various limitations exist, specifically the resource-intensive data preprocessing and inclusion of less informative features. These limitations introduce challenges when adapting the classification framework presented in C22 for our objective of identifying telescope usage (refer to Section 3 for detailed insights).

To overcome the limitations, we present a robust, interpretable, and scalable text classification framework using machine learning to automatically identify telescope usage in publications. We train and validate the text classification framework using labeled data from The Barbara A. Mikulski Archive for Space Telescopes (MAST)[8] as a case study.

In Section 2, we introduce the data used in the binary text classification task and outline the data preprocessing workflow. Additionally, we discuss the training process for the selected classifiers along with their hyperparameter tuning. In Section 3, we present the performance of the selected classifiers, utilizing various commonly used metrics, which we draw comparisons with C22. Finally, in Section 4, we draw conclusions from our findings and emphasize the effectiveness of the developed classification framework in correctly tracking the usage of MAST-hosted missions.

2. Methods

Our text classification framework is designed and trained to determine whether a publication contains the usage of any MAST-hosted mission. In collaboration with MAST scientists, we utilize manually classified data to train a supervised binary classifier, employing the labels MAST and Not MAST along with associated publication metadata in the training process. Figure 2 presents the framework, which identifies the observation sections in astrophysics publications by filtering the full text for sentences containing mission keywords (e.g., MAST and TUES). The identified observation sections are

numerically represented using term frequency–inverse document frequency (TF-IDF) for classifier training. We assess various classifiers' performance classifying the numerically transformed sentences into the two classes. Notably, the framework leverages contextual information around identified mission keywords to provide informative features for the classifier.

2.1. Data

Historically, MAST researchers manually classified astrophysics publications from NASA/ADS. Specifically, they classified 32,518 publications from 1996 to 2022, that were identified by querying NASA/ADS for publications containing one or more of the specific keywords listed in Appendix A. Publications are labeled either as MAST or Not MAST based on the presence or absence of MAST-hosted mission usage (see Appendix B for details on the classification process at MAST). To be exact, there are 10,638 MAST publications and 21,881 Not MAST publications in the database. Additionally, relevant metadata (e.g., arXiv identifier,[9] bibcode, mission, and year) is aggregated along with the label.

To compile a data set to train, validate, and test a supervised text classifier, we first obtained the full-text publications from the Semantic Scholar Open Research Corpus (S2ORC).[10] S2ORC is an open-access collection of scientific publications that allows for bulk access to full-text articles developed by the Allen Institute for AI (K. Lo et al. 2020). We extend the metadata provided by MAST scientists by leveraging the full text of the manually labeled astrophysics publications. The full text contains informative features, including the MAST-hosted mission usage description, which is not available in the metadata.

We then crossmatched the astrophysics publications available in S2ORC with the metadata from the labeled data set from MAST, utilizing the available arXiv IDs to identify the full text. The crossmatching resulted in a total of 14,808 publications in our corpus (i.e., the collection of publications) with full-text data available, comprising 7664 MAST and 7144 Not MAST publications. The balanced subset was utilized to train, validate, and test a classification model.

2.2. Tokenization and Transformation

We began our data preprocessing by splitting each publication in our corpus into smaller pieces, specifically by
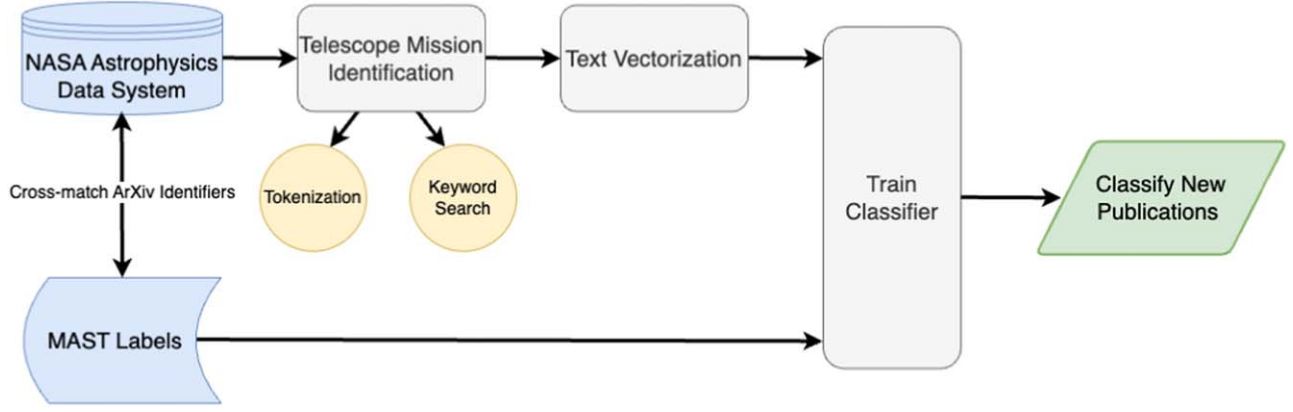

Figure 2. Our classification framework, presented here, follows a supervised training structure. First, we identify relevant sentences in each publication containing mission keywords (e.g., Kepler and TESS) and then we vectorize and classify the text.

sentence, using the sentence tokenizer from the NATURAL LANGUAGE TOOLKIT library in Python (S. Bird & E. Loper 2004). Next, we parsed the text to identify the sentences containing specific mission keywords (e.g., Kepler, K2, MAST, TESS, etc.).

We experimented with varying the sentences surrounding a mission keyword identified in the full text as input. The context varied from zero sentences either preceding or following a sentence containing an identified mission keyword to the entire full text of the publication. Table 1 presents the varied number of sentences surrounding a mission keyword utilized in training the classifiers. The context words surrounding a mission keyword included in each class differ and inform the classifier differently for the classification of publications. The preliminary testing found that exclusively utilizing the sentences containing keywords generally leads to the highest performance. Notably, the restricted context window enables the model to better distinguish between the two classes. Therefore, we chose to limit classifier training to exclusively the sentences containing mission keywords in the classification framework.

Figure 1 displays two publications filtered for a mission keyword and the context words included in the identified sentences. The classification is informed by the contextual words around a mission keyword.

We utilize TF-IDF to numerically represent the identified relevant sentences for training (H. P. Luhn 1957; K. Sparck Jones 1972). We utilize the TF-IDF formulation presented by the SCIKIT-LEARN TF-IDF vectorizer[11] (F. Pedregosa et al. 2011):

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \tag{1}$$

$$TF(t, d) = \frac{count(t, d)}{maxCount(d)} \tag{2}$$

$$IDF(t) = \log\left(\frac{1 + n}{DF(t) + 1}\right) + 1, \tag{3}$$

where t represents a term (word) in the document d across the total documents (n). The count(t, d) indicates the number of times that term t occurs in the document d, while maxCount(d) denotes the total number of terms in the document; d and DF(t) is the number of documents containing the term t across all

---

[11] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Table 1
The Table Displays the Performance of Three Classifiers (Support Vector Machine, Random Forest, and Multilayer Perceptron) as the Number of Input Sentences for Training Varies, Assessing the Required Context for Classification

# of Sentences Before, After	Support Vector Machine	Random Forest	Multilayer Perceptron
0, 0	0.929	0.905	0.908
2, 0	0.919	0.899	0.911
0, 2	0.920	0.896	0.912
2, 2	0.915	0.897	0.910
4, 2	0.904	0.888	0.904
2, 4	0.904	0.894	0.902
4, 4	0.899	0.884	0.904
Full text	0.815	0.814	0.848

Note. The Support Vector Machine classifier, trained exclusively on sentences containing mission keywords, achieves a successful accuracy of 92.9% on the test set.

documents in the set. The final TF-IDF score for each term t in each document d is given by TF-IDF(t, d).

TF-IDF uses a vocabulary to construct the numerical transformation of each publication by considering all words present across publications. To optimize our classification framework, we tuned the TF-IDF vectorizer's hyperparameters by adjusting the vocabulary size, excluding terms appearing in less than 0.001 and more than 0.999 of the publications. We also used an ngram range of (1, 2) to include two-word sequences (e.g., white dwarf), which are relevant in the domain of astrophysics. This reduced the default vocabulary from over 40,620 unique words to 29,605 words. Consequently, the chosen minimum and maximum document frequency thresholds may lead to certain rare and ubiquitous contextual words being omitted from the final TF-IDF vector representations. Each publication is transformed into a vector using the constructed vocabulary, which resulted in each TF-IDF vector representation being $1 \times 29,605$. The hyperparameter tuning reduced the computational complexity by reducing the vocabulary size and, in turn, reducing the vector representation size. TF-IDF is an interpretable representation technique as each word in a given text is directly mapped to a corresponding element in the vector representation based on the vocabulary (P. Sitikhu et al. 2019). The constructed vocabulary guides the
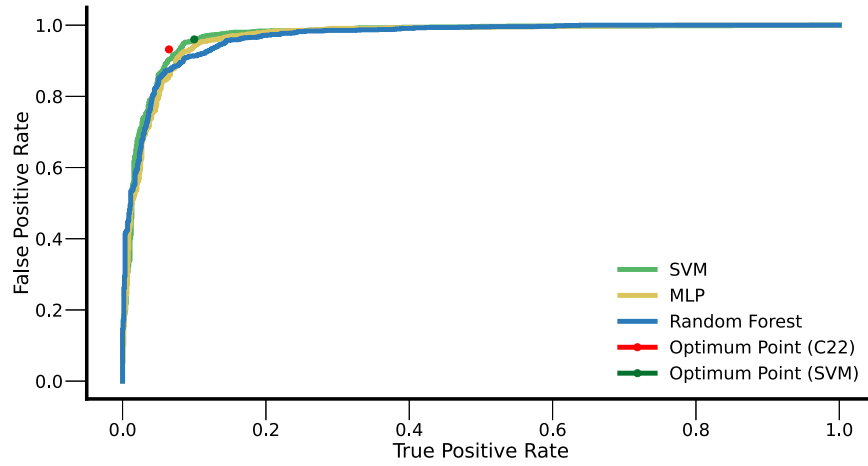
Figure 3. We present the receiver operating characteristic (ROC) curves displaying the performances of the SVM, random forest, and MLP classifiers. Notably, the SVM model trained exclusively on sentences containing mission keywords has an area under the curve (AUC) of 0.97. An AUC of 0.5 indicates that the model's ability to differentiate between the two classes is no better than random chance, while an AUC of 1 indicates the model's ability to perfectly distinguish between the two classes. The red point displays the optimum true-positive and false-positive rates from the classifier in paper C22, with a true-positive rate of 0.93 and a false-positive rate of 0.07. In comparison, the dark green point displays that our SVM model has a higher true-positive rate of 0.96. However, this improvement is accompanied by a slightly higher false-positive rate of 0.1, indicating the inclusion of some irrelevant papers.

model in recognizing informative contextual words surrounding mission keywords for classification.

2.3. Classifier Training

We chose to explore three classification algorithms: support vector machine (SVM), random forest, and multilayer perceptron (MLP), implemented in Scikit-learn. For training and evaluation, we adopted a commonly used approach by randomly splitting our data set of 14,808 publications into training, validation, and test sets. The respective ratios were set at 60%, 20%, and 20%, as determined to be close to optimal across various settings (see, e.g., K. K. Dobbin & R. M. Simon 2011).

We manually tuned the hyperparameters of the three classifiers. For the SVM, we experimented with adjusting the class weight parameters to modify the weight assigned to the MAST publications class to decrease the false-negative rate. In the case of the random forest, we configured the hyperparameters to include 500 estimators, set a verbosity level of 1, used 4 jobs, and specified a minimum of 10 samples per leaf. Similarly, we tuned the MLP utilizing an adaptive learning rate, 100 hidden layers, a ReLU activation function, and an ADAM optimizer (see, e.g., D. P. Kingma & J. Ba 2017).

There is a trade-off when adjusting the class weight parameter to decrease the false-negative rate—it will include more false positives, increasing the number of Not MAST publications classified as MAST. We prioritize minimizing the false-negative rate by maximizing the accurate classification of publications containing MAST-hosted mission usage, while retaining an acceptable false-positive rate currently set at 0.1 or lower. The threshold was determined based on its impact on the workload of MAST scientists at STScI. With the current monthly rate of approximately 550 new publications for classification and a false-positive rate of 0.1, resulting in ~55 incorrectly classified papers requiring manual review each month. The threshold of acceptable false positives is set by balancing both the classification accuracy and workload management, reducing the number of papers requiring manual

review to a level that allows MAST scientists to efficiently conduct verification while maintaining time for their research.

3. Results and Discussion

The SVM classifier, trained exclusively on sentences containing mission keywords, achieves the highest accuracy of 92.9%, up from 81.6% when trained on the full text. Throughout the rest of the paper, when mentioning the SVM model, we refer to the highest-performing SVM. In classification tasks, the common metrics used are (for more details, see B. Juba & H. S. Le 2019)

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})}.$$

The SVM classifier achieves a recall of 96.0%, a precision of 90.1%, and an accuracy of 92.9%. Meanwhile, the highest-performing random forest classifier exhibits a recall of 92.0%, a precision of 90.0%, and an accuracy of 90.5%, and the highest-performing MLP exhibits a recall of 92.0%, a precision of 92.3%, and an accuracy of 91.2%.

To provide a more comprehensive view of the three classifiers' performances, we examine the receiver operating characteristic (ROC) curve, depicted in Figure 3. An ROC curve with an area under the curve value of 0.97 signifies a high probability for the SVM model to distinguish between true positives and false positives.

Figure 4 presents the confusion matrices for the three classifiers. The SVM model demonstrates a true-positive rate of 96%, accurately predicting labeled MAST publications, and a true-negative rate of 90%, correctly identifying labeled publications as Not MAST publications. Additionally, the random forest model achieves a true-positive rate of 92% and a true-negative rate of 89%, while the MLP exhibits a true-positive rate of 92% and a true-negative rate of 92%. Manual
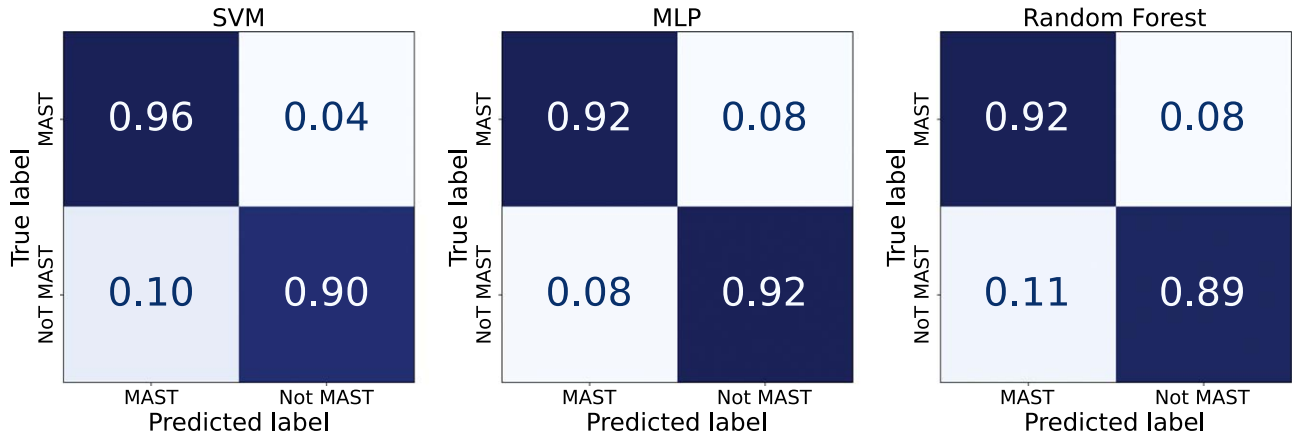
Figure 4. A confusion matrix delineates the rates of predicted and true labels, enabling the evaluation of true and false predictions for each label in our binary classification (J. Maria Navin & R. Pankaja 2016). The SVM model exhibits a true-positive rate of 96% in accurately predicting labeled MAST publications and a true-negative rate of 90% in correctly identifying labeled Not MAST publications. In comparison, both the random forest model and MLP have lower true-positive rates, while the MLP has a slightly higher true-negative rate.

classification, while more accurate, is time-consuming. The objective of employing machine learning is to automate the process and ensure comprehensive identification of publications containing mission usage. These results above underscore how well the SVM model aligns with our goal of reducing the likelihood of excluding MAST publications by minimizing the false-negative rate.

We trained the classifier presented in C22 on our data set, achieving a 90.2% accuracy. We note the limitations of comparing our work to C22, as there are differences in data, preprocessing, and the specific problem addressed with the classifier. The C22 classifier is designed for the broader needs of NED, utilizing a comprehensive full-text feature extraction method. This approach is better suited for classifying observational data across a wide range of astronomical topics, where relying on a fixed set of keywords would be insufficient. In contrast, our classifier is specifically tailored to track the usage of individual facilities, a more focused task that allows for a flexible and faster method.

Our approach, trained exclusively on sentences containing MAST-hosted mission keywords, offers several advantages. Unlike the resource-intensive feature extraction process in C22, which requires days of CPU time, our method is significantly faster and more scalable. Additionally, our classifier is less sensitive to changes in text structure or formatting, as it focuses on specific sentences in the text, compared to C22, which utilizes HTML files extracted from various astronomical journals as training data. The NED classification framework's sensitivity to text structure nuances, such as document formatting or HTML templates, necessitates periodic retraining. Particularly, the alteration of a journal's HTML template would require retraining for new publications. As the volume of publications grows exponentially, our framework scales more effectively, requiring minimal retraining, which is primarily necessary when new mission keywords are added, making it more adaptable to evolving needs (see Appendix C for performance characteristics). While the C22 classifier offers advantages for tasks across broad astronomical topics, our method provides a more targeted and efficient solution for tracking facility usage in MAST. This comparison highlights that the choice of classification method should be driven by the specific requirements of the problem at hand, with our approach being particularly well-suited for the needs of facility tracking.

4. Conclusion

We present a classification framework designed for large scientific institutions to effectively monitor the utilization of their scientific instruments and gain valuable insights into the evolving needs of the scientific community. The exponentially growing scientific literature has burdened MAST scientists who manually classify publications to track MAST-hosted mission usage, leading to a persistent backlog of unclassified publications.

Our classification framework leverages NLP techniques to automatically and accurately extract usage information. The presented framework workflow mirrors the interpretable classification methodology employed by MAST scientists, emphasizing the identification of sentences containing mission keywords within the plain text of publications. Notably, the identification of relevant sentences containing mission keywords enhances accuracy by highlighting contextual information. Considering only sentences containing mission keywords improved classification accuracy from 81.6% to 92.9%, when compared to using the full-text publication (see Appendix D for further metrics regarding full text).

The interpretability of TF-IDF enhances the framework's robustness by enabling direct evaluation of classifier performance at the word level. The explicit mapping of context words to elements of the vector representations facilitates the transparent assessment of the framework's effectiveness across various observatories. The integration of this classification model into the MAST workflow promises to alleviate manual classification and assist scientific institutions, such as STScI, in identifying the evolving needs of the global astronomical community. Future work will refine publication tracking for observatories, incorporating mission-specific and publication-type information to customize the framework for diverse observatory needs and enhance resource utilization tracking.

Contributor Roles

1. Conceptualization: Wolfgang Kerzendorf and Brian Cherinka.
2. Data Curation: Brian Cherinka and Vicente Amado Olivo.
3. Formal Analysis: Vicente Amado Olivo.
4. Funding Acquisition: Wolfgang Kerzendorf.
5. Investigation: Vicente Amado Olivo, Brian Cherinka, Wolfgang Kerzendorf, Katharina von der Wense, and Annie Didier.
6. Methodology: Vicente Amado Olivo, Wolfgang Kerzendorf, Katharina von der Wense, and Annie Didier.
7. Project Administration: Wolfgang Kerzendorf.
8. Resources: Wolfgang Kerzendorf.
9. Software: Vicente Amado Olivo.
10. Supervision: Wolfgang Kerzendorf.
11. Validation: Vicente Amado Olivo.
12. Visualization: Vicente Amado Olivo.
13. Writing—original draft: Vicente Amado Olivo and Brian Cherinka.
14. Writing—reviewing and editing: Brian Cherinka, Wolfgang Kerzendorf, Katharina von der Wense, Joshua V. Shields, and Annie Didier.

Appendix A
MAST Keywords

Full list of keywords used by MAST Scientists to search for candidate MAST publications.

1. PAN STARRS	22. AIDA
2. PANSTARRS	23. BEFS
3. PS1	24. Copernicus
4. Pan-STARRS	25. EUVE
5. Pan-STARRS-1	26. Extreme Ultraviolet
6. Pan-STARRS1	27. FUSE
7. PanSTARRS-1	28. Far Ultraviolet
8. PanSTARRS1	29. GALEX
9. WUPPE	30. Galaxy Evolution
10. UIT	31. HUT
11. TESS	32. Hopkins Ultraviolet
12. TUES	33. IMAPS
13. Berkeley Extreme and Far-UV Spectrometer	34. IUE
14. Berkeley Spectrometer	35. 10.17909
15. Transiting Exoplanet Survey Satellite	36. K2
16. Tubingen Ultraviolet Echelle Spectrometer	37. KTWOCANDELS
17. Ultraviolet Imaging Telescope	38. Kepler
18. International Ultraviolet Explorer	39. MAST
19. Interstellar Medium Absorption Profile Spectrograph	40. OAO-3
20. Wisconsin Ultraviolet Photo-Polari-meter Experiment	41. ORFEUS
21. Orbiting Retrievable Far and Extreme Ultraviolet Spectrometers	

Appendix B
MAST Manual Classification Process

In the following, we describe the manual classification process utilized by the MAST scientists.

B.1. Keyword Search

The MAST archive scientists classify only a subset of papers from NASA/ADS. First, a full-text keyword search is done for all papers that contain any mission or MAST-related keywords, e.g., MAST, Kepler, and TESS (see the full list in Appendix A). Flagship missions like the Hubble Space Telescope Mission and James Webb Space Telescope Mission are separately classified. This search is done periodically, aggregated by month, currently producing, on average, roughly 550 publication results each month for further human verification.

B.2. Classification Criteria

Next, the publications are classified and labeled with the following information:

1. *Mission*: the relevant mission name (i.e., TESS, Kepler).
2. *Paper type*: which paper category the publication fits into (e.g., a "science paper" or a "mention").
3. *Ignored*: whether the publication should be ignored altogether.

Publications marked as Ignored are typically papers where the matched keyword refers to something unrelated to astronomical mission data at MAST. For example, the MAST acronym keyword can reference "Mega Amp Spherical tokamak," rather than "Mikulski Archive for Space Telescopes," or the K2 mission keyword can reference a variable in a mathematical equation, k^2, rather than NASA's K2 mission. The contaminations in the initial keyword search are due to limitations with the NASA/ADS keyword search and must be accounted for during classification. See the Not MAST box in Figure 1 for an additional example of a keyword mismatch.

For binary classification, the archive scientists created a new column indicating for each paper whether it is a MAST paper or not. Any paper with at least a mention of a relevant keyword is considered a MAST paper. All papers that were classified as "ignores" are considered as Not MAST papers.

Appendix C
Performance of Classifiers

C.1. Code Availability

The code and data are open-source and are available here[12] as a Zenodo repository.

C.2. Profiling

The performance of our classification framework was profiled for execution time utilizing Python's timing functionality, with all measurements repeated over seven runs to calculate mean times and standard deviations. We profiled the best-performing SVM, random forest, and MLP classifiers. The data processing steps, comprising tokenization, keyword search, and TF-IDF computation, exhibited the following execution times per run: tokenization required 4 minutes 8 s (±694 ms), keyword search took 12.3 s (±56.7 ms), and TF-IDF computation lasted 5.5 s (±27.7 ms).

The random forest classifier trains in 11.7 s (±47.4 ms) per run with an accuracy of 90.5% on the test set. In comparison, the MLP trains in 210 s (±29.6 s) with an accuracy of 90.8%. The SVM, as discussed in Section 3, is the classifier used in the framework as the accuracy of 92.9% is the highest of the tested classifiers, while the training time is 40.3 s (±218 ms) per run.

---

[12] doi:10.5281/zenodo.14014142.

Appendix D
SVM Full-text Metrics

In Figures 5 and 6 we plot the ROC curve and confusion matrix for an SVM model trained on the full-text publications.
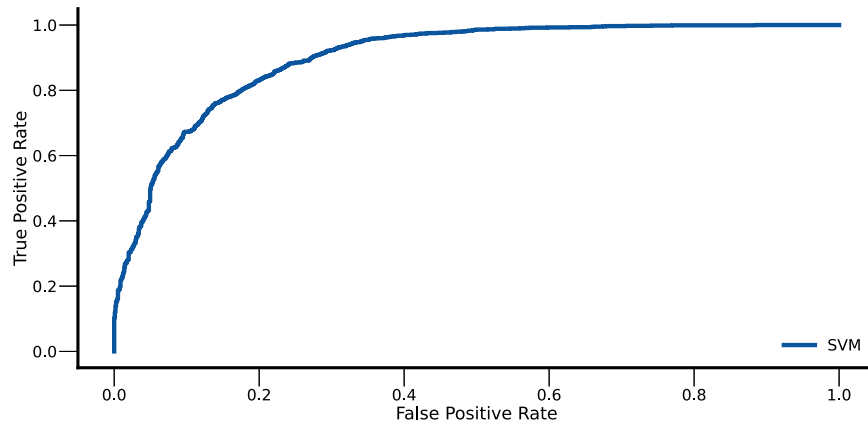

Figure 5. The ROC curve of the SVM model trained on the full-text publications has an area under the curve (AUC) of 0.90 compared to 0.97 when identifying the sections of the publication containing keywords.
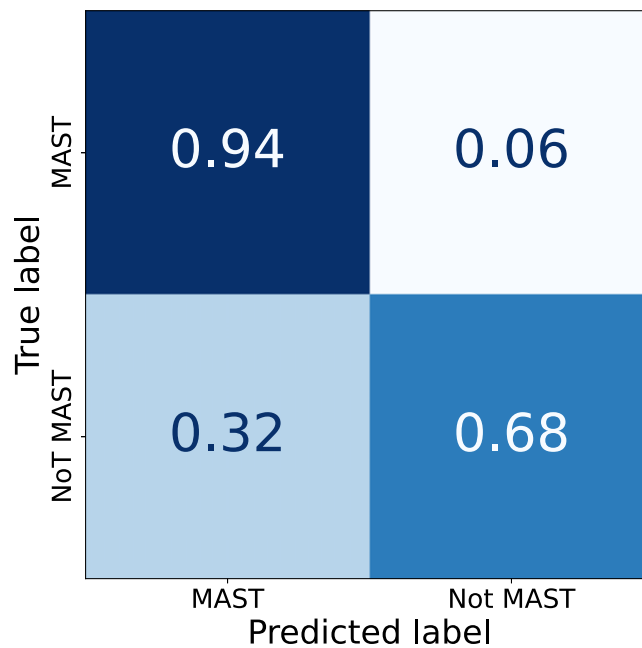

Figure 6. The SVM model using the full-text publications, instead of identifying the sections with keywords, exhibits a true-positive rate of 94% in accurately predicting labeled MAST publications and a true-negative rate of 68% in correctly identifying labeled Not MAST publications.

ORCID iDs

Vicente Amado Olivo ⬤ https://orcid.org/0000-0003-2248-0941
Wolfgang Kerzendorf ⬤ https://orcid.org/0000-0002-0479-7235
Brian Cherinka ⬤ https://orcid.org/0000-0002-4289-7923
Joshua V. Shields ⬤ https://orcid.org/0000-0002-1560-5286

References

Barnett, T., Godjevac, S., Renders, J.-M., et al. 2009, DESI III: The ICAIL Workshop on Globaal E-Discovery/E-Disclosure (Princeton, NJ: Citeseer)

Beltagy, I., Lo, K., & Cohan, A. 2019, in Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), ed. K. Inui et al. (Cambridge, MA: MIT Press), 3615

Bird, S., & Loper, E. 2004, in Proc. ACL Interactive Poster and Demonstration Sessions (Cambridge, MA: MIT Press), 214, https://aclanthology.org/P04-3031

Bornmann, L., Haunschild, R., & Mutz, R. 2021, Humanit. Soc. Sci. Commun., 8, 1

Chen, T. X., Ebert, R., Mazzarella, J. M., et al. 2022, PASP, 134, 014501

Dobbin, K. K., & Simon, R. M. 2011, BMC Med. Genomics, 4, 31

Farshid, D., & Dastani, M. 2023, Digit. Health, 9, 20552076231185674

Grothkopf, U., Meakins, S., Bordelon, D., et al. 2018, EPJ Web Conf., 186, 06001

Juba, B., & Le, H. S. 2019, in Proc. AAAI Conf. on Artificial Intelligence, Vol. 33 (Washington, DC: AAAI), 4039

Kerzendorf, W. E. 2019, JApA, 40, 23

Khadhraoui, M., Bellaaj, H., Ammar, M. B., Hamam, H., & Jmaiel, M. 2022, ApSci, 12, 2891

Kingma, D. P., & Ba, J. 2015, arXiv:1412.6980

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. 2020, in Proc. 58th Annual Meeting of the Association for Computational Linguistics (Cambridge, MA: MIT Press), 4969

Luhn, H. P. 1957, IBMJ, 1, 309

Maria Navin, J., & Pankaja, R. 2016, IJETR, 6, 75

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. 2019, in 2019 Artificial Intelligence for Transforming Business and Society (AITB) (New York: IEEE), https://ieeexplore.ieee.org/document/8947433

Sparck Jones, K. 1972, J. Doc., 28, 11